Apprentissage de données fonctionnelles par modèles multitâches et application à la détection pour le sport de haut niveau

Arthur Leroy

Résumé:

Ce travail porte sur l'analyse de données fonctionnelles et la définition de modèles multitâches pour la régression et la classification non supervisée de telles données. Une application d'aide à la décision pour la détection de jeunes sportifs à fort potentiel pour le sport de haut niveau est également proposée, et sert de fil rouge pour l'illustration pratique des méthodes et algorithmes développés dans cette thèse. Nous nous intéressons particulièrement à l'étude de séries temporelles, souvent observés à pas de temps irréguliers, issues de multiples individus, supposés partager de l'information commune. Ces données sont souvent étudiées comme des observations ponctuelles de processus fonctionnels sous-jacents, et l'on peut voir ce problème comme l'étude d'une collection de courbes. La méthode centrale de la thèse, et l'algorithme d'apprentissage qui lui est associé, s'intéresse à la régression fonctionnelle multitâche à l'aide d'un modèle mixte de processus Gaussiens. Ce cadre probabiliste non paramètrique permet de définir une loi a priori sur la fonction d'apprentissage reliant les données d'entrée et de sortie. Le calcul explicite de la loi a posteriori du processus Gaussien à l'aide d'un échantillon d'observation fournit une prédiction probabiliste (moyenne du processus et incertitude via la variance) en tout point non observé de l'espace des entrées (pour tout temps dans le cas de séries temporelles). Nous proposons une extension multitâche à cette méthode classique d'apprentissage supervisé en définissant un modèle mixte, où une donnée fonctionelle est supposée être la somme d'un processus Gaussien moyen, commun à tous les individus, et d'un processus Gaussien spécifique à l'individu. L'inférence est effectuée à l'aide d'un algorithme EM pour estimer les hyperparamètres du modèle et la loi a posteriori du processus moyen. De nouveaux calculs de loi a posteriori permettent ensuite de fournir une prédiction, pour un nouvel individu, qui est améliorée par l'utilisation de l'information commune avec les autres individus à travers le processus moyen. Cette prédiction intègre l'incertitude à la fois commune et individuelle à travers la covariance de la loi a posteriori. Cette méthode montre une nette amélioration des performances prédictives comparée aux alternatives et étend le cadre d'application aux séries observées irrégulièrement (nombre et position des temps d'observations différents). Une extension de ce modèle est ensuite proposée à travers un modèle de mélange de processus Gaussiens. A l'aide d'un algorithme variationel EM associé, il est possible d'effectuer un clustering des individus en parallèle de l'apprentissage des autres quantités d'intérêt. Cette approche permet d'étendre l'hypothèse d'un processus moyen sous-jacent à plusieurs clusters ayant chacun leur processus moyen. Les prédictions ainsi obtenues sont à présent cluster-spécifiques, tout en conservant les propriétés précédemment évoquées. Après la mise en évidence, dans un premier travail, de structures de groupes au sein des données réelles sportives, les méthodes décrites ci-dessus ont été utilisées pour fournir, à travers leurs prédictions probabilistes, un outil d'aide pour la détection des jeunes sportifs prometteurs pour le haut niveau.

Mots clefs : Régression fonctionnelle, clustering de courbes, processus Gaussiens, apprentissage multitâche, algorithmes EM, ...